Using Microarray Data to Inform and Identify Breast Cancer Therapy Targets

Ethan Ensminger

University of Minnesota, Twin Cities

April 30, 2015

1 Introduction

More than 220,000 women every year are diagnosed with breast cancer, making it the second leading cause of cancer death in women [1]. Most of these deaths are caused by cancer that has spread to other parts of the body. Indeed, 30% of women who are diagnosed while in the early stages of breast cancer will have the cancer recur or have metastatic disease. Furthermore, greater than 70% of patients diagnosed have invasive tumors [2].

In fighting this disease, many approaches have been used to try to understand and specifically target breast cancer. One technique that has been particularly powerful for gaining a better perspective of the underlying mechanisms driving disease has been microarray technologies. Microarray data show expression levels of genes within a population of cells at a specific time point. This data is analyzed quantitatively and statistically to find abnormal levels of genes which can be caused by misregulation of cancerous cells. In some cases, this misregulation helps drive disease progression.

One particular study has provided microarray data from 295 breast cancer patients in addition to extensive clinical data such as the following: presence of metastatic tumors, tumor size, and patient survival [3, 4]. This dataset is publicly available and is the dataset used for this project.

By identifying genes that predict and correlate with clinical outcomes, and particularly individual genes in specific pathways displaying coordinated behavior, we hope to identify novel therapeutic candidate targets.

2 Data

The microarray and clinical data for this project was downloaded from `http://changlab.stanford.edu/2005-PNAS-Data.html`. From this url, two files were used. `NKI_Expression_data_complete.txt` contains the microarray expression data (log10 format) and `Clinical_Data_Supplement.xls` contains the clinical data for the 295 patients that participated in the study.

2.1 Microarray data preparation

As always, the data needed to be prepared for further analysis. Duplicate columns in the middle of the microarray data set were removed. These were probably left over from copy-and-pasting when the original file

was made. Contig barcodes were mapped to Genbank EST accession numbers using the file `ArrayNomenclature_contig_accession.xls` from `http://bioinformatics.nki.nl/data/van-t-Veer_Nature_2002/`. Finally, the name of the columns were changed by concatenating the gene name (if available) with its systemic name (e.g. Genbank accession number). Because multiple probes can map to the same gene, this gives each column a unique name. Alternatively, each duplicate gene column could be combined into a single aggregate column of expression levels. However, this was not done for this analysis.

Since our analysis is based on [5], all expression levels were converted to z-scores.

2.2 Adding gene names using BLAST

The original data is from 2005. Many of the probes do not have a gene name associated with them (11,121 probes). Since most pathway analysis tools use gene names rather than Genbank accession numbers, it was decided to run the accession numbers with no gene name through BLAST to see if any more gene names could be added. A standalone version of BLAST (version 2.2.30+) was used along with a copy of the rna sequence files. `blastn` was performed on accession numbers in the dataset that did not have a gene name associated with it.

At the command line, the following command was used for each of the query files (query files contain accession numbers to align):

```
blastn -query NoName_query01.txt -db refseq_rna -out NoName_output01.txt -outfmt "6 qacc
    sacc sscinames stitle pident evalue"
```

Since BLAST returns multiple sequence alignments for most queries, the alignment with the lowest e-value was chosen (e-value is a kind of p-value that also takes into account sequence length and is calculated by the BLAST software). If multiple alignments have the same e-value, the one with the highest coverage was chosen.

After adding the BLAST results, probes with gene names increased from 13,360 to 18,919. However, 5562 still have no gene name.

2.3 Caching p-values

For our initial analysis, many p-values needed to be calculated. Since there are 24,481 probes, it was decided to cache the results in order to save computing time.

On looking at the data, it seemed reasonable to truncate the fraction of patients in the high expression group between 0.1 and 0.9 to make the log-rank test valid (see Figure 3). This means that at 0.1, about 30 patients will be in the high expression group, while at 0.9, about 30 patients will be in the low expression. Between 0.1 and 0.9, 100 log-rank tests were performed for each gene. This translates to 100 p-values per gene.

2.4 Software for analysis

The Python programming language was used for data preparation and analysis. Major packages used were Pandas (a package for easily manipulating tabular data) and lifelines (used for Kaplan-Meier curves and log-rank testing). Jupyter (formerly known as IPython) was used as the programming environment.

Preliminary pathway analysis was done using the Reactome database (`reactome.org`).

3 Analysis and results

Briefly, we want to use the microarray data and clinical data to find significant genes present in breast cancer. We can combine the microarray data and clinical data by doing multiple log-rank tests to find significant p-values. These p-values are used to rank genes. By changing three variables (a p-value threshold, minimum number of "hits", and consecutive hits), we can produce lists of genes that can be run through pathway analysis software. The results of the pathway analysis can show what pathways are affected by the genes in the lists produced, which in turn could lead to new drug targets.

3.1 Background

The initial idea for analyzing the microarray data in the way described later in this article came from [5]. In this paper, they presented a survival curve like Figure 1, which showed the percentage of patients that lived metastasis free. The patients were divided into two groups, depending on whether they expressed higher or lower amounts of FAK1. In this particular paper, two-thirds were in the high expression group. The distribution of z-scores is presented in Figure 2 and shows where the patients were partitioned for the Kaplan-Meier curve in Figure 1.

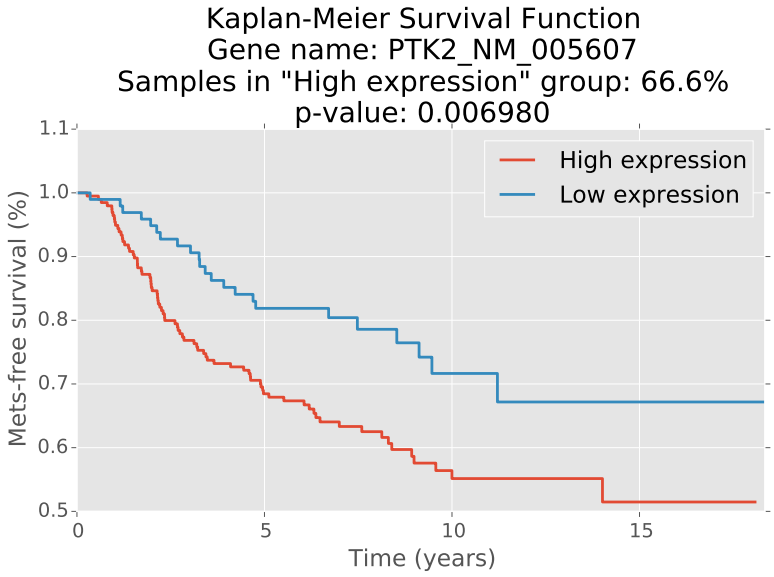

Figure 1: Reproducing figure from [5] showing the survival curve for FAK1 (or PTK2). P-value was calculated using the log-rank test.

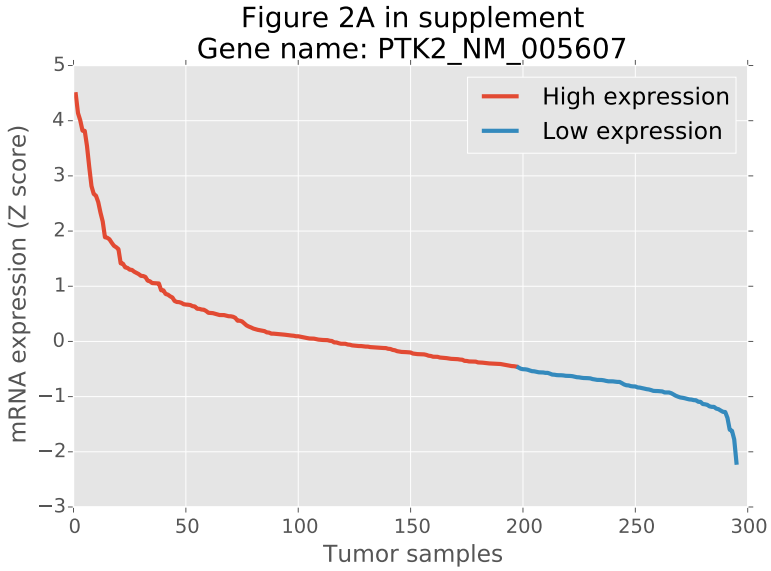

Figure 2: From [5] showing the distribution of z-scores across 295 patient samples. Data sorted by z-score.

3.2 Plotting multiple p-values

It seemed that partitioning the patients at 2/3 was somewhat arbitrary. To see how p-value varied with the partition location, it was decided to start with 10% (or about 30) of the patients in the high expression group and incrementally add patients to the high expression group until 90% of the patients were in the high expression group. A log-rank test was performed after every transfer. Figure 3 shows how p-value changes as more samples are shifted to the high expression group. Figure 4 is the same as Figure 3 except that the maximum value on the y-axis has been lowered to 0.05. Keep in mind that each point in these two plots is the p-value of a single log-rank test and that the partition between groups moved so that one group gains about 2 patients and the other loses 2 patients.

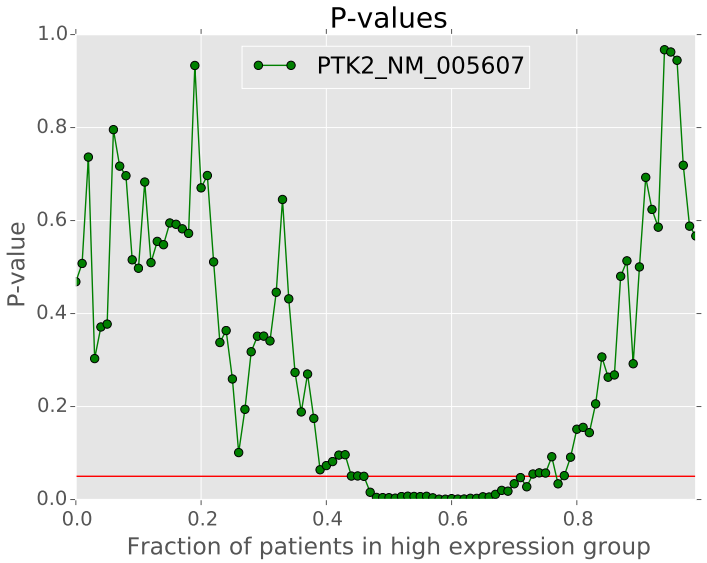

Figure 3: Plots how the p-value from log-rank tests varies with the partitioning of high and low expression groups. Red line indicates a p-value of 0.05.

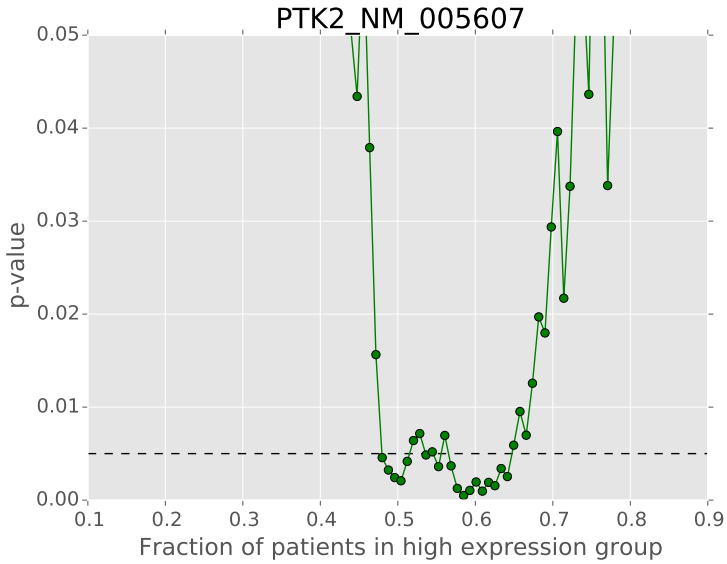

Figure 4: Same as Figure 3, but with the upper limit of the y-axis lowered from 1 to 0.05. The black, dashed line indicates a p-value of 0.005.

Referring to Figure 3, a common trend was that the p-values at the ends would rise sharply. A possible explanation is not having enough patients in one of the groups. This is the reason the analysis was restricted to looking at p-values where the fraction of patients in the high expression group varied between 0.1 and 0.9.

Although the high p-values occurred in many of the genes, some genes did not show this behavior. A possible reason is that some patients express very high levels or very low levels of a gene (see Figure 2). This causes one of the groups to be unfairly weighted and the log-rank test to conclude that the two groups are significant, even though only a few patients are in one of the groups.

Another hypothesis is that these extreme patients are in fact a subgroup of the total data. One could possibly define a set partition at some value like 0.7 and see whether certain genes have similar numbers of hits. However, enough patients need to be in the groups to give statistically valid results.

Focusing on the middle portion of Figure 3, there is a region of low p-values. The length of this region could indicate how robust the partition is in that region. For example, if the region were only two or three points, one could transfer 4 or 5 patients and the two groups would not be significant. Defining stability as the number of consecutive hits will be used later as one of two ways to create gene lists.

3.3 Using p-values as a ranking criterion

As seen from Figure 4, depending on where the partition between high and low expression is set, the p-value will change. In order to use these p-values as ranking criteria, the p-values needed to be aggregated. There are many possible ways to do this, but we decided to do two things: First, set a p-value threshold. Because lower p-values are understood to be more significant, log-rank tests with p-values less than or equal to this threshold are labeled as a hit. Second, we count the number of hits for each gene. In this way, genes with more hits are assumed to be more significant. Therefore, the number of hits can be used to rank genes.

Figure 5 shows how the number of significant genes vary with p-value threshold and minimum number of hits. The most important conclusion is that the number of genes is affected more by the p-value threshold than by the minimum number of hits.

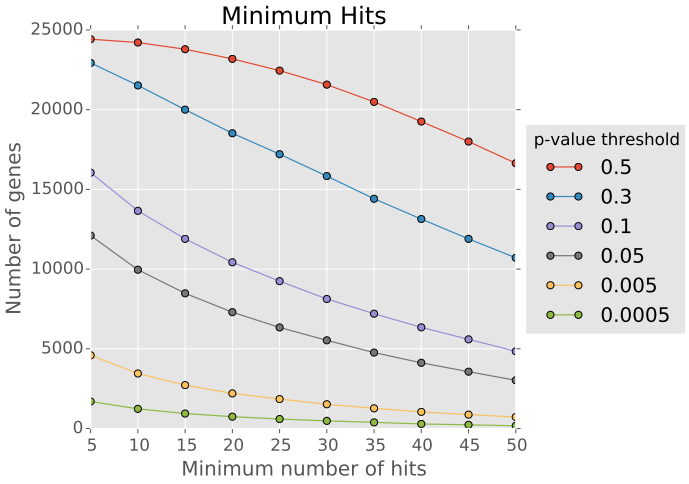

Figure 5: A plot of the number of genes that have at least a certain number of hits below the p-value threshold. Maximum possible hits is 100.

Another way to generate gene lists was hinted in the previous section. Most genes have places on the p-value plots where the p-value stays very low over an extended region (Figure 3). This can be used as measure of how robust that gene is and can be used to generate different gene lists. Figure 6 shows how the number of genes varies with p-value, but this time as a function of how wide the stable region is, which is the number of consecutive hits below the p-value threshold. As in Figure 5, the number of genes is more strongly affected by p-value threshold. Also, using consecutive hits, the number of genes drops more rapidly as the minimum number of consecutive hits is increase.

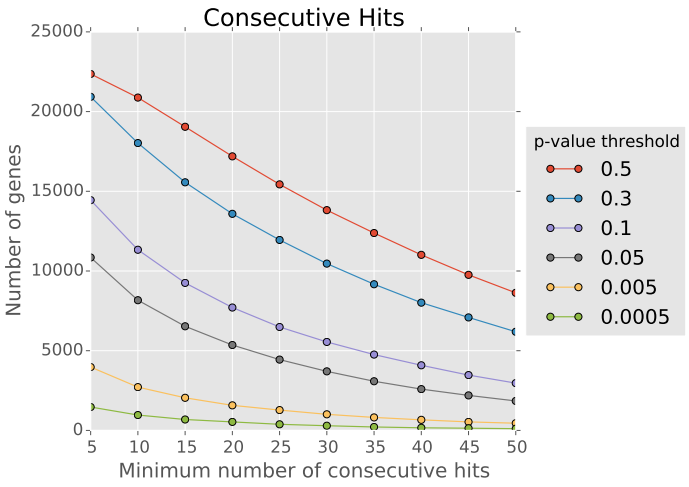

Figure 6: Illustrates how the number of valid genes varies with p-value threshold and consecutive hits. Maximum possible hits is 100.

3.4 Creating gene lists

To feed into pathway analysis tools, gene lists needed to be generated. As a first pass, the parameters were solely based on having a gene list of manageable size, but still contain as many genes as possible. Based on Figure 5, a p-value threshold of 0.005 and a minimum number of hits of 20 were chosen. The distribution of the list generated is shown in Figure 7. Table 1 shows the top ten hits for the p-value threshold of 0.005 and minimum hits of 20. After removing duplicate gene names, the total number of genes in the list was 2138.

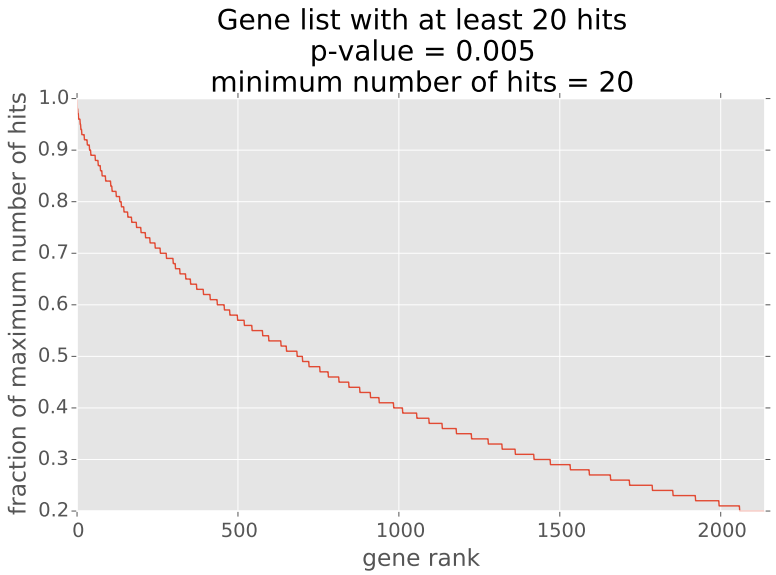

Figure 7: Showing the number of hits for each gene. Maximum number of hits is 100. Duplicate gene names were removed. Total number of genes is 2138.

Table 1: The first 10 genes with a p-value threshold of 0.005 and minimum of 20 hits.

gene_name	hits
BIRC5	100
ARHI	98
DKFZP586E1519	98
UBCH10	97
DKFZp762E1312	97
PRC1	96
AW137640	96
STK15	96
SESN3	96
KIAA0165	96

A second gene list using consecutive hits was also generated. The distribution of hits is plotted in Figure 8 and the top 10 hits are shown in Table 2. The total number of genes with consecutive hits of at least 20 was 1528 (after removing duplicate gene names).

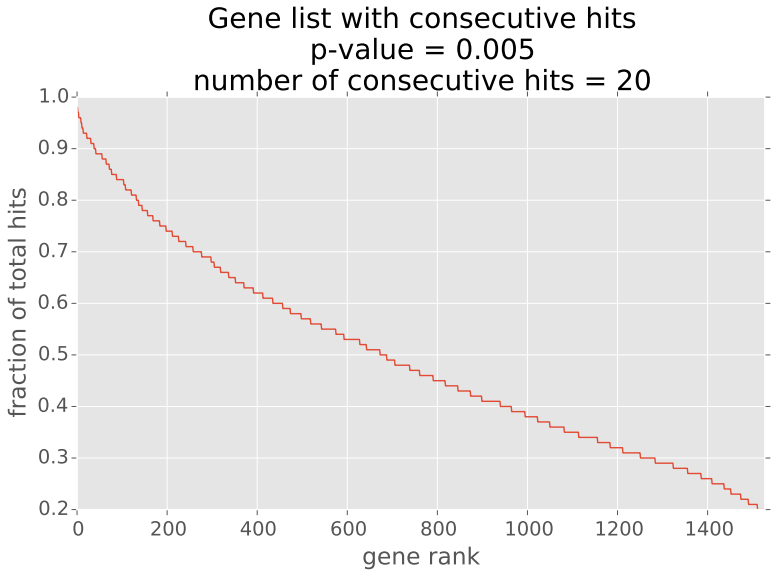

Figure 8: The distribution of hits using consecutive hits.

Table 2: The first 10 genes using a p-value threshold of 0.005 and consecutive hits of at least 20

gene_name	hits
DKFZP586E1519	98
ARHI	98
DKFZp762E1312	97
UBCH10	97
PRC1	96
SESN3	96
AW137640	96
KIAA0165	96
STK15	96
PSMD7	95

3.5 Preliminary pathway analysis

Table 3 shows the first 10 pathways for minimum number hits of 20 and p-value of 0.005 using `reactome.org`. The results are sorted by p-value for that pathway. In total, 1257 pathways were identified using this gene list. Unfortunately, 1338 gene names in the list were not found in the database. This means about 62.6% of the genes in the list were not included in the pathway analysis.

Table 3: Top ten pathway results for minimum hits of 20. Duplicate gene names were removed before submitting to Reactome.org.

Pathway name	# Entities found	Entities pValue	Entities FDR
Nonsense Mediated Decay (NMD) independent of t...	41	0.000143	0.105741
Peptide chain elongation	40	0.000183	0.105741
Eukaryotic Translation Termination	39	0.000434	0.124386
Eukaryotic Translation Elongation	40	0.000504	0.124386
Formation of a pool of free 40S subunits	38	0.001667	0.267658
Nonsense-Mediated Decay (NMD)	43	0.001705	0.267658
Nonsense Mediated Decay (NMD) enhanced by the ...	43	0.001705	0.267658
G1/S-Specific Transcription	11	0.002328	0.305014
SRP-dependent cotranslational protein targetin...	42	0.002480	0.305014
G2/M Checkpoints	23	0.003476	0.386538

For the gene list with consecutive hits of 20 and p-value of 0.005, the top 10 pathways are presented in Table 4. Total number of pathways identified was 1127. As with the minimum hits gene list, about 943 genes were not found in the pathway database. This means that about 61.7% of the gene names were not used in the pathway analysis.

Table 4: Top ten pathway results for consecutive hits of 20. Duplicate gene names were removed before submitting to Reactome.org.

Pathway name	# Entities found	Entities pValue	Entities FDR
Nonsense Mediated Decay (NMD) independent of t...	33	0.000072	0.038668
Peptide chain elongation	32	0.000109	0.038668
Eukaryotic Translation Termination	32	0.000132	0.038668
Eukaryotic Translation Elongation	32	0.000271	0.056625
Cell Cycle, Mitotic	113	0.000357	0.058220
Cell Cycle Checkpoints	38	0.000477	0.058220
Nonsense-Mediated Decay (NMD)	35	0.000509	0.058220
Nonsense Mediated Decay (NMD) enhanced by the ...	35	0.000509	0.058220
Polo-like kinase mediated events	12	0.000540	0.058220
Formation of a pool of free 40S subunits	31	0.000560	0.058220

4 Discussion

4.1 Improving BLAST results

The first improvement is completely mapping accession numbers to gene names. As stated in the data section, 5562 probes do not have an associated gene name. It turns out that there were problems while doing the BLAST search. Apparently, BLAST can only handle about 500 accession numbers at once. This may be limited by the memory capacity on the computer used. The end result was that not all accessions sent to BLAST were actually submitted to BLAST.

4.2 Cutoffs for making gene lists

For this project, the p-value threshold, minimum hits, and minimum consecutive hits were set somewhat arbitrarily. The p-value threshold of 0.005 was set based solely on keeping the gene list to a manageable size. Now that the first run has been completed, this p-value threshold will be set more appropriately. We can look at p-value profile for genes that are already known to play important roles in breast cancer. For instance, FAK is known to have a significant impact [6]. Looking at Figure 4, we already see that we should probably change the threshold to at least 0.01 in order to include FAK.

The minimum number of hits is not as important as setting the p-value threshold because the number of genes is affected more by the p-value threshold (see Figure 5). The same trend holds for consecutive hits. The minimum number of hits and consecutive hits parameters will likely be set based on studying the results from pathway analysis, since these parameters affect the genes that are at the bottom of the generated list of relevant pathways. Also, there is a bug in the filter used to for consecutive hits. If you compare Table 1 with Table 2, BIRC5 should also be in Table 2. However, this does not affect the results very much because that is the only gene missing from Table 2.

4.3 Preliminary pathway analysis

From Tables 3 and 4, the top results for both gene lists are more or less the same. But it is interesting that the p-values are lower for the consecutive hit list. This could possibly happen because the gene list for minimum hits is larger (800 gene names) than the gene list for consecutive hits (585 gene names). However, the p-value could also be lowered because the consecutive hits list is more specific, making the pathways found more significant.

In contrast to the top of both pathway lists, the list of pathways become increasingly diverse. As mentioned in the previous subsection, the minimum hits or consecutive hits limits will affect these more diverse pathways. More pathway analysis and possible mRNA knockdown experiments will be needed to verify a relevant limit.

As stated earlier, approximately 60% of the gene names in both lists were not found in the Reactome database. This is partially due to not having complete BLAST results. Even after correcting the BLAST problem, a few gene names will still not be filled because some “accession numbers” are actually antibody barcodes. However, on looking at the list of gene names not found, many are actual gene names. A possible explanation is that the Reactome database is not large enough. Further analysis with other pathway analysis tools, such as Ingenuity Pathway Analysis (IPA, `http://www.ingenuity.com/products/ipa`), may ameliorate this problem.

5 Conclusion

By performing log-rank tests, we were able to develop a unique ranking criterion based on the p-values for multiple partitions of the 295 patients. Using this criterion, we ranked genes. We also developed a measure of stable regions of p-values by looking at the number of consecutive hits. Using these parameters (p-value threshold, minimum consecutive hits, or minimum hits), we created lists of genes that are possibly significant with respect to metastasis-free survival. These gene lists were run through the Reactome pathway database and some preliminary analysis was done. In addition, we also have ideas for extending our work such as doing more pathway analysis, RNA knockdown experiments, and using other clinical data included in the dataset. We hope that the process of creating these gene lists and subsequent pathway analysis will lead us to new drug targets and to new pathways to target in the fight against breast cancer.

Acknowledgments

I would like to thank Dr. Paolo Provenzano for being willing to mentor this project and taking time to discuss results. I would also like to thank the Undergraduate Research Opportunities Program (UROP) for supporting this project.

References

[1] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics 2015. *CA: A Cancer Journal for Clinicians*, 65(1):5–29, Jan 2015.

[2] ACS. Cancer Facts and Figures. 2012.

[3] Marc J. van de Vijver, Yudong D. He, Laura J. van 't Veer, Hongyue Dai, Augustinus A.M. Hart, Dorien W. Voskuil, George J. Schreiber, Johannes L. Peterse, Chris Roberts, Matthew J. Marton, Mark Parrish, Douwe Atsma, Anke Witteveen, Annuska Glas, Leonie Delahaye, Tony van der Velde, Harry Bartelink, Sjoerd Rodenhuis, Emiel T. Rutgers, Stephen H. Friend, and René Bernards. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *New England Journal of Medicine*, 347(25):1999–2009, Dec 2002.

[4] H. Y. Chang, D. S. A. Nuyten, J. B. Sneddon, T. Hastie, R. Tibshirani, T. Sorlie, H. Dai, Y. D. He, L. J. van't Veer, H. Bartelink, M. van de Rijn, P. O. Brown, and M. J. van de Vijver. From The Cover: Robustness scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proceedings of the National Academy of Sciences*, 102(10):3738–3743, Feb 2005.

[5] Y. Pylayeva, K. M. Gillen, W. Gerald, H. E. Beggs, L. F. Reichardt, and F. G. Giancotti. Ras- and PI3K-dependent breast tumorigenesis in mice and humans requires focal adhesion kinase signaling. *J. Clin. Invest.*, 119(2):252–266, Feb 2009.

[6] Florian J. Sulzmaier, Christine Jean, and David D. Schlaepfer. FAK in cancer: mechanistic findings and clinical applications. *Nat Rev Cancer*, 14(9):598–610, Aug 2014.